# Visualizing Complexity

Amy Braverman
*Jet Propulsion Laboratory*

## 1. Introduction

Complexity is an important feature of geophysical data for two reasons. First, geophysical data sets are often quite large, and for exploratory purposes it is useful to know where the data are easily understood, and where more effort will be required to understand their content. Second, data complexity often mirrors complexity of the underlying geophysical processes, where more elaborate physical models may apply.

The connection between the problem of summarizing large data sets and the quantization problem in signal processing provides a framework within which to define and analyze data set complexity. In signal processing the objective is to transmit random signals over a channel with limited capacity. Signals can't transmitted exactly because an infinite number of bits would be required to specify each possible signal value, also called a realization. Instead, possible signal values are grouped, or quantized, into clusters, and when a particular signal is encountered, an integer identifying its cluster is transmitted. At the receiver, the cluster identifier is replaced by a representative value for the group as an estimate of the original signal. Often cluster representatives are the cluster mean realizations. The number of clusters, the distribution of realizations to them, and the probabilities of encountering each realization determine the average number of bits required to specify group membership for a random signal:

$$h = -\sum_{k=1}^{K} p_k \log p_k, \qquad (1)$$

where $p_k = N_k/N$ is the proportion of possible signal values belonging to the $k$th cluster, $N_k$ is the number of possible signal values belonging to the $k$th cluster, $N = \sum_{k=1}^{K} N_k$, and $K$ is the number of clusters. $h$ is the entropy of the quantization. Of course, this procedure also creates error between the original signal and its representation at the receiver. On any single transmission the error can be measured by $\|x - \hat{x}\|^2$ where $x$ is the original signal and $\hat{x}$ is the cluster representative for $x$. The average error over all possible realizations is called distortion:

$$\delta = \frac{1}{N}\sum_{n=1}^{N} \|x - \hat{x}\|^2. \tag{2}$$

In signal processing the objective is to find a quantization scheme that achieves minimum distortion among all those with entropy less than or equal to channel capacity. The quantization paradigm is also useful in a data analytic setting: we regard data points in a data set as signal realizations, and use quantization to cluster them into groups. Then the set of group means and populations is a compact summary of the original data. Moreover, if we use the quantizer with minimum entropy among all schemes with distortion less than or equal to a fixed level, that entropy can be regarded as a measure of data complexity at that distortion level. The two situations embody complementary constrained optimization problems that can be solved using the same algorithm: Entropy-constrained Vector Quantization (ECVQ; Chou, Lookabaugh, and Gray, 1989).

Various modifications and generalizations of ECVQ have been used to find good quantizers in signal processing (Gersho and Gray, 1992, de Garrido and Pearlman, 1995) and in data analysis (Braverman, 2001). In the latter case ECVQ is modified to cope with large data volumes often encountered in geoscience. These additional, practical constraints slightly increase entropy over what would have been obtained using the original version of ECVQ. In practice, however, they are not too far above the unmodified ECVQ values, and provide a useful estimate of data set complexity for purposes of exploratory data analysis.

**To visualize complexity in large, spatial, geophysical data sets, we partition the data on a spatial grid, say 1° × 1°, and treat data belonging to each grid cell as a separate data set. ECVQ modified for data analysis is used to quantize each data set independently. A map of the resulting entropies shows regions of high and low data complexity.** The method is demonstrated here using test data from the International Satellite Cloud Climatology Project (ISCCP; Rossow, Walker, Beuschel, and Reuter, 1996). These data are described in the next section.

## 2. ISCCP Test Data

Here we focus on one month's worth of ISCCP pixel-level (DX) data. These data were obtained from NASA's Langley Atmospheric Sciences Data Center. They include sufficient information to determine whether each 30 km pixel is clear or cloudy, and for cloudy pixels, cloud-top pressure and optical thickness are reported. For our test data we use measurements of cloud-top pressure and optical thickness for cloudy pixels in January, 1993. Each data point is a column vector, $\boldsymbol{y} = (y_1, y_2)'$, where $y_1$ is cloud-top pressure and $y_2$ is optical thickness.

There are approximately 40 million such data points in January 1993, and these are partitioned into 1° × 1° data sets. Let $\boldsymbol{y}_{t,\theta,\phi}$ denote a data point acquired at time $t$, latitude $\theta$, and longitude $\phi$. Then the set of all $\boldsymbol{y}_{t,\theta,\phi}$ with $u \leq \theta < u+1$ and $v \leq \phi < v+1$ are denoted by $[u, v]$. We write $[u, v]$ as a matrix with $N_{u,v}$ rows and two columns. The rows are $\boldsymbol{y}'_{t,\theta,\phi}$'s belonging to the 1° × 1° spatial region with southwest corner at $(u, v)$. Note that time is ignored in this formulation. Data points in a 1° × 1° data set are indexed by $n$ so that $\boldsymbol{y}_{n,u,v}$ is the $n$th data point in $[u, v]$. Figure 1 is a map showing grid cell populations, $N_{u,v}$, for the test data.

We seek to characterize the complexity of all $[u, v]$, $u = -90, \ldots, +89$ and $v = -180, \ldots, +179$. This is accomplished by applying ECVQ modified for data analysis to each $[u, v]$ independently. ECVQ and its modification for data analysis are described in the next section.

## 3. Algorithm

ECVQ (Chou, Lookabaugh, and Gray, 1989) is an iterative descent algorithm for finding minimum entropy quantizers subject to a constraint on distortion, or equivalently, minimum distortion quantizers subject to a constraint on entropy. Operationally it accomplishes this by minimizing the Lagrangian objective function,

$$L(\lambda) = \frac{1}{N} \left[ \sum_{n=1}^{N} \|\boldsymbol{y}_n - \beta[\alpha(\boldsymbol{y}_n)]\|^2 + \lambda \left( -\sum_{n=1}^{N} \log N[\alpha(\boldsymbol{y}_n)] \right) \right].$$

$\lambda$ is a Lagrange multiplier, $1[\cdot] = 1$ if its argument is true and zero otherwise, and $N = \sum_{k=1}^{K} N(k)$. $\beta$ and $\alpha$ are defined as follows. Let $\mathcal{I} = 1, 2, \ldots, K$ be a set of integers called the index set. $\alpha : \mathcal{R}^2 \rightarrow \mathcal{I}$ accepts a data point as its argument, and returns an integer indexing the group to which $\boldsymbol{y}$ is assigned. $\beta : \mathcal{I} \rightarrow \mathcal{R}^2$ accepts a group index, and returns the corresponding group representative:

$$\beta(k) = \frac{1}{N(k)} \sum_{n=1}^{N} \boldsymbol{y}_n 1[\alpha(\boldsymbol{y}_n) = k], \quad \text{where} \quad N(k) = \sum_{n=1}^{N} 1[\alpha(\boldsymbol{y}_n) = k].$$

A quantizer for $\boldsymbol{y}$ is $q(\boldsymbol{y}) = \beta[\alpha(\boldsymbol{y})]$ .

The basic ECVQ algorithm is:

1. Fix $K$, the initial number of groups, and $\lambda$, the Lagrange multiplier.
2. Randomly assign each $\boldsymbol{y}_n$ to one of the $K$ clusters by specifying initial values for $\alpha(\boldsymbol{y}_n)$. Compute initial group representatives and populations, $\{\beta(k), N(k)\}_{k=1}^{K}$. Then reassign each $\boldsymbol{y}_n$ to the group with the nearest euclidean distance $\beta(k)$, and update $\{\beta(k), N(k)\}_{k=1}^{K}$..
3. Set $\alpha(\boldsymbol{y}_n)$ to the group index for which the penalized distance $d_\lambda(\boldsymbol{y}_n, k) = \|\boldsymbol{y}_n - \beta[\alpha(\boldsymbol{y}_n)]\|^2 + \lambda\left(-\log N[\alpha(\boldsymbol{y}_n)]\right)$ is minimized. Delete empty groups, and update representatives and populations.
4. Repeat the previous step until the algorithm converges, which it is guaranteed to do in a finite number of iterations.

Note first that one does not specify the desired level of distortion or entropy for the constraint directly, but rather a value of $\lambda$. $\lambda$ can be interpreted as specifying the importance of entropy relative to distortion in the objective function. Second, in step 3 $\boldsymbol{y}_n$ may be assigned to a group not nearest to it in Euclidean distance if a sufficiently more populous group exists further away. $\lambda$ controls the magnitude of this effect. Some of the original groups may become empty and if so they are removed from further consideration. Assignment of data points to groups under ECVQ will be more concentrated than had $d(\boldsymbol{y}_n, k)$ been defined as ordinary Euclidean distance. At termination, $\alpha$ provides a mapping of data points to groups, and a representative and associated population for each group. The final number of groups, $\tilde{K}$, may be fewer than the initial number. A diagram of ECVQ is shown in Figure 2a.

**In order that ECVQ be practical for data analysis, it is modified in three ways.** First, ECVQ is performed only on a sample of data points. Once the algorithm produces a set of assignments, associated group representatives, now called preliminary representatives, are used to form new groups of original data points. That is, each original data point is assigned to the nearest sample-derived representative in Euclidean distance. The means and populations of these final groups are reported to summarize the data, and the associated entropies used to assess data set complexity. This modified form of ECVQ is called Extended ECVQ (EECVQ), and the reason for using it is that ECVQ is iterative and too intensive to run on large data volumes. Using samples reduces computational burden. An additional benefit is that the final assignment step creates an approximately nearest neighbor mapping of data points to groups. This is not the case at termination of ECVQ, since ECVQ minimizes $L(\lambda)$, not distortion. EECVQ summaries have minimum mean squared error as estimates of the data they summarize. However, they are not, strictly speaking, minimum entropy.

The second modification is to perform ECVQ or EECVQ multiple times on a given data set, each time using a different initial assignment of data points to groups. This is necessary because algorithm solution depends on the initial random assignments. To increase chances of finding good solutions, we run with a variety of starting configurations, and choose the solution that achieves the minimum value of the appropriate loss function: $L(\lambda)$ for ECVQ or final distortion for EECVQ. In the case of EECVQ each sample automatically provides a different starting configuration because each contains a different set of sampled data points. We call this procedure Monte Carlo ECVQ or EECVQ (MCECVQ or MCEECVQ). An additional benefit of the Monte Carlo approach is that it allows us to assess the effect of randomness. By looking at the distribution of the loss function values we get an idea of how stable the algorithm is for the data to which it is applied.

Finally, in some cases the data may be so large that scanning them once per run to do the final assignments for MCEECVQ may not be feasible. In that case, a third modification is invoked. On each run, an independent sample of data points is used in place of the original data for the final assignment step. The distortion obtained by grouping this sample is an estimate of the true final distortion that would be obtained grouping the whole data set. After all runs, the set of preliminary representatives yielding the smallest final distortion estimate is used to group the entire, original data set. This modification requires just two complete passes through all the data: one to collect the required samples and another to group the original data. This version of the algorithm is the one used in the next section to assess complexity in the ISCCP data, and is summarized in Figure 2b.

## 4. Application to ISCCP Test Data

We ran MCEECVQ with the third modification on each 1° × 1° spatial subset of the ISCCP test data independently. Data were first standardized using the global means and standard deviations of cloud-top pressure and optical thickness computed from all the data. We used $K = 30$ initial groups, sample sizes of 300, 30 runs, and a common value of $\lambda = .04$. $\lambda$ was determined by experimenting on a subset of grid cells (those with latitudes and longitudes both evenly divisible by 5°) to find the value tending to equalize the final distortions as much as possible across grid cells. This criterion was used because we desire a collection of grid cell summaries which are of approximately equal quality. Otherwise, differences in the summaries, and in particular in their entropies, might reflect differences in the quality of their fits to their data rather than differences in the data sets. The 53,087 populated grid cells each took about half a minute to process on a combination of 250 MHz RISC 10000 and 400 MHz RISC 12000 SGI Origin 2000 processors. The grid cell data summaries have from one to 30 groups. Figure 3 is a map showing the number of groups used to summarize each grid cell, and Figure 4 is a map of the entropies associated with each spatial region.

There is considerable spatial continuity in Figure 4, and certain regions corresponding to geophysical phenomena are readily identifiable. The intertropical convergence zone (ITCZ) appears as a band of relatively high complexity along the equator in the eastern Pacific ocean. There is a large region of low complexity data just to the south and expanding along the coast of South America. The Bermuda high, monsoonal moisture in the southwestern United States are also visible.

# 5. Summary and Conclusion

The algorithm described here provides one way of quantifying and visualizing data complexity in large spatial data sets. Data are partitioned on a spatial grid, and a measure of complexity derived for the data in each spatial region. The quantitative definition of complexity used here is an information-theoretic one: the average number of bits required to specify group membership of data points under the optimal grouping scheme. The optimal grouping is the one that minimizes the average number of bits required among all groupings with distortion not exceeding a specified level. Therefore, complexity is theoretically measured at some level of distortion, and complexity of data in one spatial grid cell relative to another may change as that distortion level varies. Here we do not set the distortion level explicitly, but rather select a value for the Lagrange multiplier in the algorithm's objective function that specifies the relative importance of distortion and entropy in deriving the grouping scheme. That value is used for all grid cells and selected to equalize grid cell distortions as much as possible so the resulting complexities are as comparable as possible. Other compromises to accommodate large data volume render complexities produced by this algorithm estimates of the true values. There is reason to believe the estimates are slightly biased above, but since we are interested in making relative comparisons in an exploratory setting, this is not thought to be a serious problem.

Two points are worth making in conclusion. First, complexity says nothing about data content, it only says something about how easily data are described. A logical way of using complexity information might be to create a set of reduced-volume proxy data sets for further analysis. Data in simple $1^\circ \times 1^\circ$ grid cells could be replaced by their means, while data in more complicated cells could be replaced by various other quantities such as maximums, minimums, or quantiles of interest. One could thus construct a variety of global maps representing different scenarios under which to test models, and thereby assess the impact of data variability. Complexity can be used to guide in this process. Finally, at least some areas having common levels of complexity in Figure 4 could've been identified on physical grounds alone. For example, the areas which comprise the ITCZ are subject to a common set of conditions and forces. It's reasonable that data collected there would have similar characteristics including complexity. The real questions are if and how more complex physical processes come to manifest themselves in more complex data, and whether areas of unexpected complexity point to previously unidentified processes of interest. These are subjects of on-going research.

# References

Braverman, A., (2001) "Compressing Massive Geophysical Data Sets Using Quantization", *Journal of Computational and Graphical Statistics*, to appear.

Chou, P., Lookabaugh, T., and Gray, R., (1989) "Entropy-constrained Vector Quantization", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, pp. 31–42.

Garrido, D.P. and Pearlman, W. (1995) "Conditional Entropy-Constrained Vector Quantization: High-Rate Theory and Design Algorithms", IEEE Transactions on Information Theory, Volume 41, Number 4, pp. 901–916.

Gersho, A., and Gray, R.M., (1992) Vector Quantization and Signal Compression, Kluwer Academic Publishers, Norwell, MA.

Figure 1. Grid cell populations, $N_{u,v}$, January 1993.

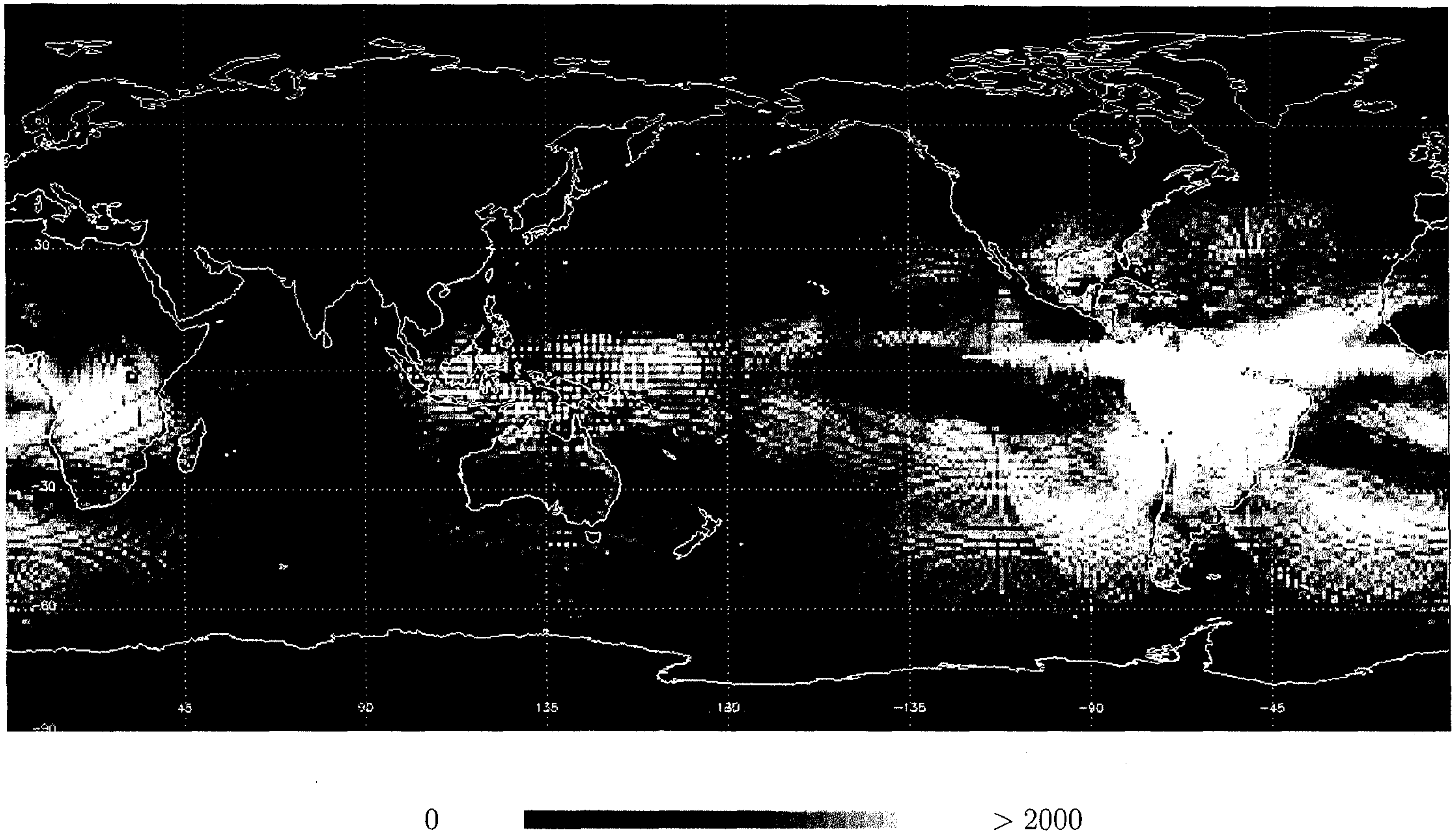


Figure 2a. ECVQ algorithm.

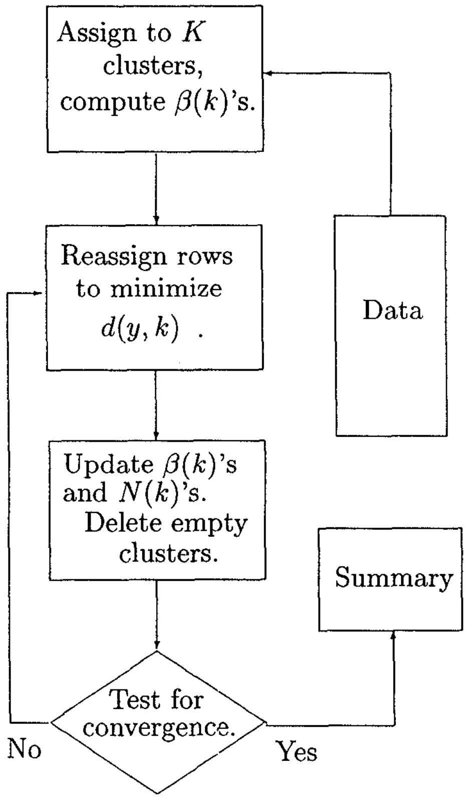

Assign to $K$ clusters, compute $\beta(k)$'s.
Data
Reassign rows to minimize $d(y,k)$ .
Update $\beta(k)$'s and $N(k)$'s. Delete empty clusters.
Summary
Test for convergence.
No
Yes

Figure 2b. MCEECVQ algorithm.

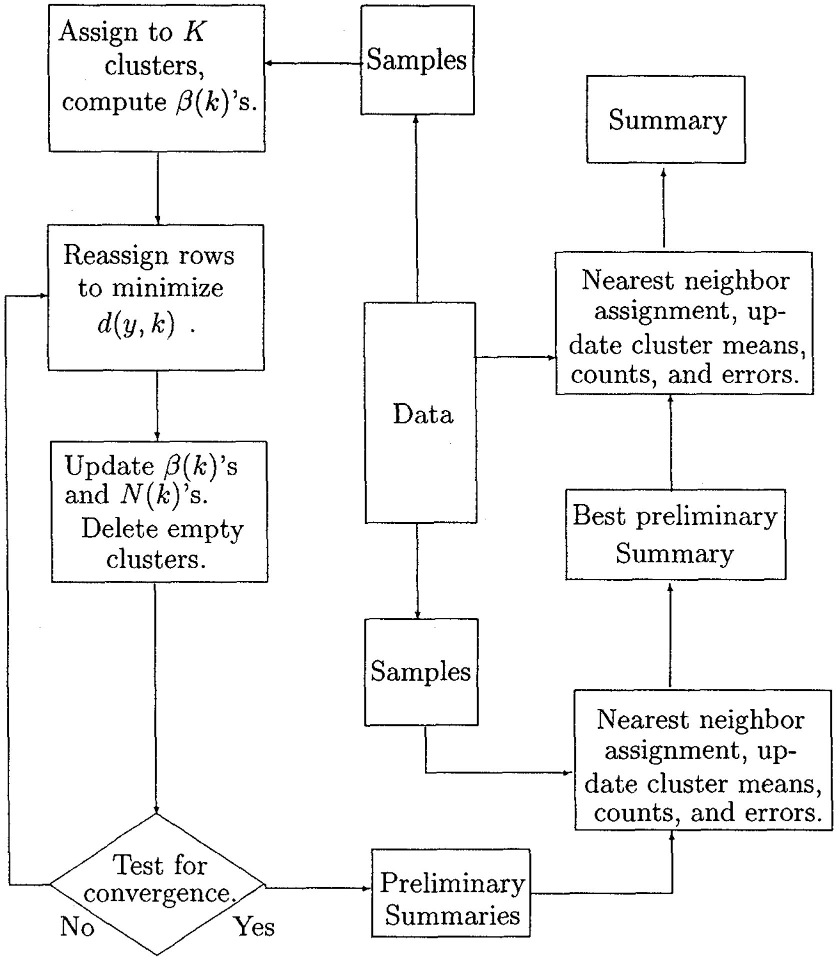
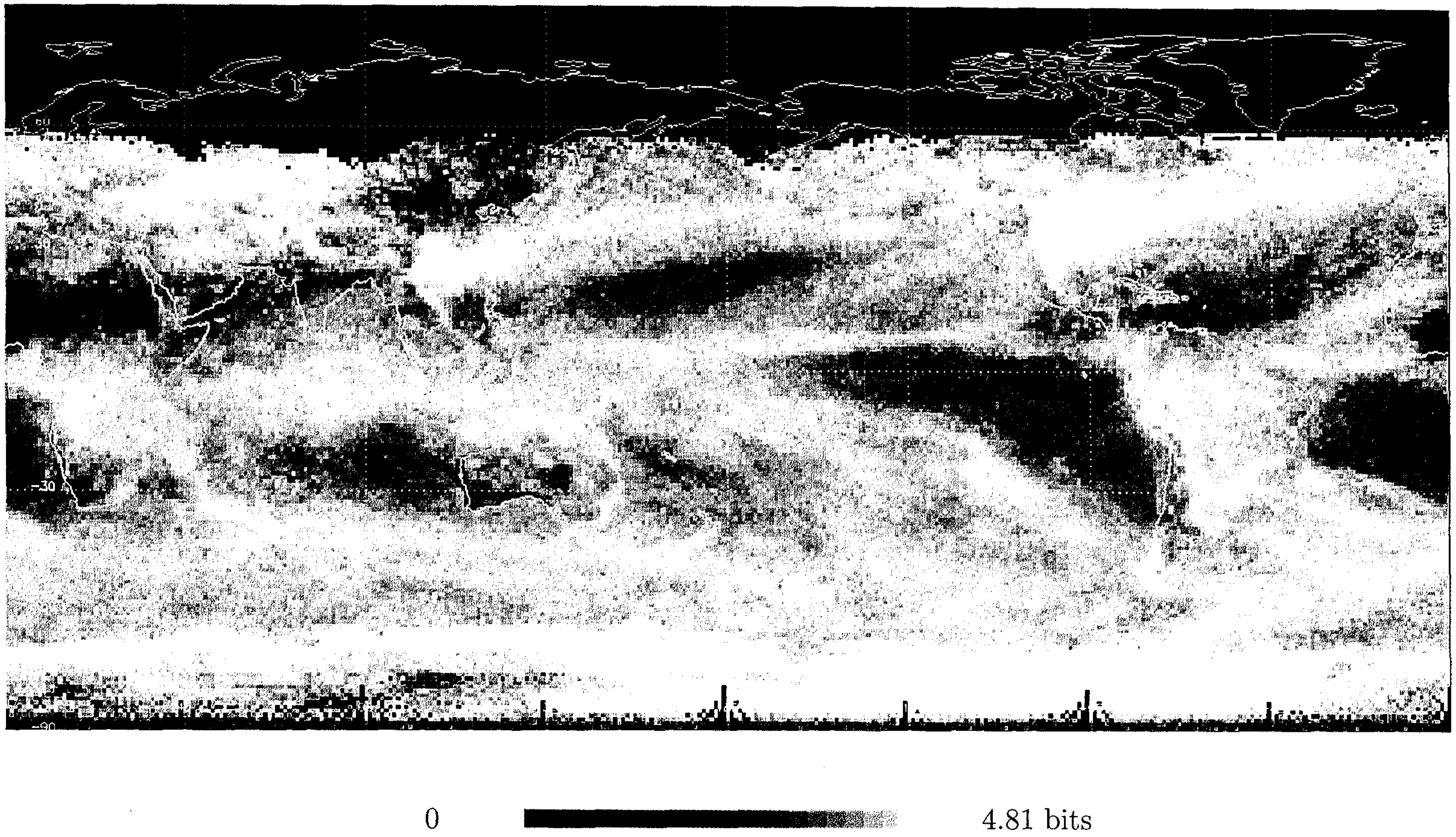

Assign to $K$ clusters, compute $\beta(k)$'s.
Samples
Summary
Reassign rows to minimize $d(y,k)$ .
Data
Nearest neighbor assignment, update cluster means, counts, and errors.
Update $\beta(k)$'s and $N(k)$'s. Delete empty clusters.
Best preliminary Summary
Samples
Nearest neighbor assignment, update cluster means, counts, and errors.
Test for convergence.
No
Yes
Preliminary Summaries

Figure 1. Grid cell entropies, $h_{u,v}$, January 1993.


60
-30
-90
0
4.81 bits